Usability of Online Grocery Systems: a focus on errors

Mark Freeman **Alison Norris** **Peter Hyland**
School of Information Technology and Computer Science
University of Wollongong
Wollongong, New South Wales
{mbf02, aen01, phyland}@uow.edu.au

ABSTRACT
Recording of errors in regards to the usability of systems has traditionally focused on safety-critical systems and business support systems. This study applies Zapf et al.'s 'Taxonomy of Errors' to a non-work related context, an Online Grocery System. The taxonomy was found to show that similar types of errors were made by all users of such systems. However, the number of errors that were recorded by different user groups varied. This finding was in contrast to previous studies, and supported the common perception that beginner users make a greater number of errors than more experienced users.

Author Keywords
Human Error, Usability, Taxonomy of Errors

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces, H.1.2 User/Machine Systems.

INTRODUCTION

The consideration of human error is of critical importance when assessing the usability of online systems. To date, there has been no research conducted on the types of errors that are made by users of online systems that are used for a non-work related activity, such as an Online Grocery System. The goal of this study is to compare the errors that are made by participants with that of the 'Taxonomy of Errors' (Zapf et al. 1992). This study sought to identify, classify and compare the errors made by different types of users on the interface of an Online Grocery System. This process allowed the 'Taxonomy of Errors' to be evaluated for its applicability to non-work related systems used by casual users.

LITERATURE REVIEW

The definition for human error used in this study is taken from Reason (Reason 1990 p. 9): "Error will be taken as a generic term to encompass all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency." While the term 'human error' is often used to "describe a certain class of human actions" (Hoc et al. 1995 p. 9), this definition of human error is restrictive. The term 'human error' can refer to both the cause of an error, and the class of actions. It has been suggested that the term 'erroneous action' is more appropriate (Rauterberg and Felix 1996) as it clearly defines an event that was at a later time deemed to be incorrect. When considering the complexity of the observable behaviour, Rauterberg and Felix (1996) assume that human behaviour by itself cannot be erroneous. Errors occur based on the decisions that humans make and the behavioural consequences of these decisions.

The predominant research on errors in computer-based systems to date has focused on 'safety-critical systems', such as air-traffic control systems (Leadbetter et al. 2001). A limited amount of research has been conducted on errors in office situations, with the results used in the development of the 'Taxonomy of Errors' (Zapf et al. 1992; Brodbeck et al. 1993). The application of the 'Taxonomy of Errors' to other domains has been limited to statistical-based analysis software (Hyland 2001) and student use with the statistical package SPSS (Harper et al. 2004). This study will demonstrate use of this taxonomy with an Online Grocery System – a system that is used voluntarily and is outside the workplace environment.

An understanding of human behaviour and the integral role of errors in human behaviour is essential prior to the assessment of such errors. Three levels of human behaviour can be identified using a classification that refers to the errors in that behaviour (Wickens and Hollands 2000): error-free skilled behaviour (classified as normal behaviour); inefficient behaviour (where there is a simpler procedure to complete the same task); and erroneous behaviour (where the result of the task does not achieve the desired outcome). Rasmussen (1986) states "if a system performs less satisfactorily than it normally does – because of a human act – the cause will very likely be identified as a human error." Psychological research into human errors determined that humans learn from unsuccessful behaviour - that is they learn from making mistakes. This understanding is critical when considering the general western attitude towards errors, which suggests that errors are something that must be avoided (Rauterberg and Felix 1996). Research suggests that the behaviour of novice users during interaction with a system is usually more complex than experts' behaviour (Rauterberg 1996, p. 828), and includes more errors.

An understanding of systems is needed to reduce the number of erroneous actions occurring. Although erroneous actions are unavoidable they are not necessarily harmful (Hoc et al. 1995), and even advanced users can encounter situations where their actions may be erroneous (Hoc et al. 1995). Erroneous actions are no longer seen as

OZCHI 2006, November 20-24, 2006, Sydney, Australia.
Copyright the author(s) and CHISIG
Additional copies are available at the ACM Digital Library (http://portal.acm.org/dl.cfm) or ordered from the CHISIG secretary (secretary@chisig.org)
OZCHI 2006 Proceedings ISBN: 1-59593-545-2

a weakness of human cognition but as a result of a mismatch or a discrepancy between the users and the system. This is similar to how Zapf et al. (1992) describe errors for their taxonomy.

TAXONOMY OF ERRORS

Sutcliffe and Rugg's (1998) taxonomy of error types and Zapf et al.'s 'Taxonomy of Errors' (1992) are both based on Reason's (1990) assessment of cognitive errors. However, the nature of the errors assessed by the two taxonomies are significantly different. Sutcliffe and Rugg's taxonomy (1998) addresses errors related to safety-critical systems, where further analysis of the errors is required to prevent future similar errors. The 'Taxonomy of Errors' (Zapf et al. 1992) identifies office workplace errors which do not lead to harm or loss in human life. The 'Taxonomy of Errors' can be applied to a range of computer-based activities, and is used in this research.

The following section will describe in full the Zapf et al. (1992) 'Taxonomy of Errors' which is divided into three different types of erroneous actions: usability problems; inefficiencies and functionality problems. It is generally assumed that novices make more erroneous decisions than experts (Zapf et al. 1992). However, Zapf's application of the 'Taxonomy of Errors' determined that this assumption was not accurate. The research also identified that different user groups make different types of errors, as classified by the 'Taxonomy of Errors'.

The first classification of erroneous actions is that of usability problems. These occur when goals or plans are not accurately developed, and signals used to guide the user are overlooked. A large number of usability problems are "knowledge errors" which are usually performed by novice users. There are four types of knowledge errors, three of which are similar to error types proposed by Hacker in 1986 (see Zapf et al. 1992) and Rasmussen (1986). These are Flexible Action Patterns (well known sub-plans are not executed or well known feedback is ignored); Sensorimotor (mistakes in using the keyboard or pointing devices); Knowledge Base (the user doesn't know a command, special key or rule used in the system); and Intellectual Regulation (a complex, conscious plan of action is developed to meet the goal but it is inadequately developed or parts of it are forgotten) (Zapf et al. 1992).

The second classification of erroneous actions is that of inefficiency problems. Inefficiency problems can result from a lack of knowledge in the actions that can be performed by a system, such as when a user performs a round about method for completing a task as they do not know a better way. Experience with a system is likely to reduce the incidence of such problems, therefore Zapf et al. (1992) suggest that "Novices should be more inefficient than experts in the knowledge subcategory of inefficiency". Inefficiency problems can also be due to the habits of a user. In this instance, the user understands that there is another, preferred way to perform a task, however due to their habits, a longer method which requires more steps is used. This type of inefficiency problem usually occurs with experts as they typically have a better understanding of systems, and have developed their preferred method of use.

The third classification of erroneous actions is that of functionality problems. These problems are typically out of the control of the user and can occur with both hardware and software. There are four types of functionality problems: action blockage (user has to give up or change goal); action repetition (a part of one's work is lost and needs to be redone); action interruption (the user is interrupted by the system, but is able (usually after some additional work) to continue); and action detour (the user knows the weaknesses of the software and compensates them).

The 'Taxonomy of Errors' has been subjected to a number of construct validations (Prümper et al. 1992; Brodbeck et al. 1993; Hyland 2001) involving participants with a variety of skill levels. However, previous studies have not included non-work related tasks conducted by casual users, such as using an Internet-reliant grocery system to purchase groceries online. Since these studies did not contain casual or novice users, it is possible that some types of errors may not have been observed. This will be achieved as part of this study

A diagrammatic overview of The 'Taxonomy of Errors' is presented in Figure 1.

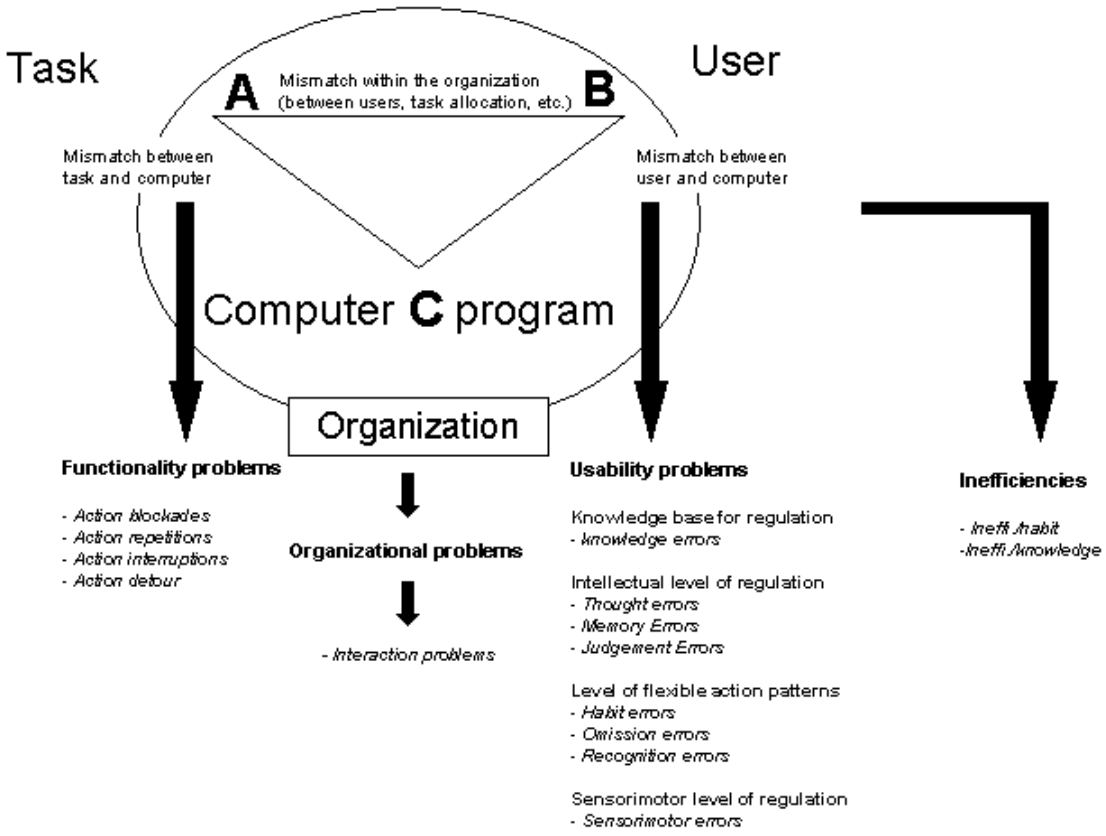

Figure 1. Zapf's Taxonomy of Errors

METHODOLOGY

Usability testing was the chosen Usability Evaluation Method (UEM) for observing the interaction of participants with the chosen website. It is based on scientific research (Rubin 1994), where controlled experiments are conducted and the outcomes recorded. These results are then examined by a tool (in this study, SPSS) to identify trends in the data. On completion of the usability testing, participants completed a post-test questionnaire which recorded participant perceptions of the Online Grocery System. The concept of the testing methodology is taken from the classical approach for conducting a controlled experiment (Rubin 1994).

The recommended usability testing group sample size varies amongst the experts in the field, and is dependent on the type of study that is being conducted. The recommendations for qualitative testing range five (Nielsen 2000) and six users (Kalin 1999), up to 10-12 users as the minimum needed (Rubin 1994). For the use of quantitative tests, Nielsen recommends testing be conducted with 20 users. The discrepancies between experts indicate that there is no agreed size for usability testing groups. For this study a sample size of 54 participants has been chosen, consisting of three types of users: beginner, intermediate and advanced users of e-commerce websites. Each of these groups consists of 18 participants. This is based on statistical calculations by Cochran and Cox (1957, p. 24).

Usability testing participants were selected on a number of criteria, to ensure a representative sample of participants that use or are likely to use online grocery websites. The level of prior experience in an online shopping environment (not limited to Online Grocery Systems), including the type and number of items previously purchased online, was used to categorise users into one of three user groups: beginner, intermediate, or advanced. Participants ranged in age, education and Internet experience. Many of the participants were university students, as people in this demographic have been previously recognised as the most likely to purchase items online (Ahuja et al. 2003). This is especially true for online grocery shopping in the future. The university students that participated in this research represented a range of subject areas and included undergraduate and postgraduate students.

Through the application of stringent controls, each participant's experience was as similar as possible. The setup of the testing room remained unchanged throughout the testing, with all testing conducted on the same

computer. All participants were given the same briefing on the testing, based on a prepared script. While all participants used the same website, minor changes made by the company were unavoidable (such as when a product becomes unavailable at the warehouse and is removed from the system). All participants were given the same set of printed materials for the test. These materials were: the background/screening questionnaire (to gather demographic information and to act as the screening process for participants); the tutorial exercise to be conducted; the Shopping Lists for each activity (task-based usability testing on an Australian Online Grocery System); and a post-test questionnaire for each Shopping List. Three Shopping Lists were completed. Each user completed List 1 (Short Test: 10 items, single quantity) and List 2 (Medium Test: 20 items, varied quantity) in a single session. List 3 (Long Test: 50 items, varied quantity) was completed in a second session conducted approximately one week later.

Participants had the entire process explained to them, and were instructed to carry out the tutorial. Once the tutorial was completed, Camtasia Studio Recorder recording software was activated to record all actions on the computer. All mouse movements, mouse clicks, screen changes and text input were captured by Camtasia Studio Recorder in an avi file. This software also recorded the time for each event. Participants completed the Shopping List(s) by purchasing the products in any order and using any means that they chose. Participants then completed the post-test questionnaire taking into account the Shopping List they had just finished.

The success of participants in completing the Shopping Lists (usability testing tasks) was assessed through the review of the Camtasia Studio Recorder footage based on a number of review criteria. This process was based on the premise that, to gain information about usability, the first rule is to watch the participants work with the website, and not just listen to what they say (Nielsen 2001). The elements of the footage that were reviewed are:

- The method that the participant employed to add an item or multiples of an item to the trolley;
- The time that it took for participants to add each item to their trolley;
- The errors that the participant made whilst adding the item(s) to the trolley;
- The total length of time and the overall navigation process employed by the participant.

For the purposes of this study, any interaction with the system that did not lead to the direct completion of adding an item to the shopping trolley was considered to be an error. This is consistent with the 'Taxonomy of Error' usage as described by Zapf et al. (1992). The time taken to recover from an error was also recorded.

One of the primary objectives of the usability testing was to record the strategies used by participants to complete the Shopping Lists. These strategies can be identified by the use of the search method or the linking method to find a product. For each product, there could be multiple attempts to find the product and each of these attempts was recorded. Note was taken of the method used to successfully add each product to the trolley and to select the quantity.

ANALYSIS OF ERRORS

The analysis of errors was initially conducted based on the total number of errors made and time taken to recover from the errors during the three lists. After this raw analysis of the errors, a detailed analysis was conducted using the 'Taxonomy of Errors' (Zapf et al. 1992). Error-handling time is "defined as the time it takes to correct or abandon correcting an error after it has been detected" (Prümper et al. 1992). As some errors were never detected by participants, no time could be associated with these types of errors. Errors were timed to the nearest second using Camtasia Studio's progress clock on the footage.

The total number of errors made by participants across all three tests was 4705. Advanced participants made 947 errors; intermediate participants made 1661 errors; and beginner participants made 2097 errors. A distribution of the total number of errors is shown in Figure 1. The mean for all participants for all tests was 87.1 with a standard deviation of 49.46 and a median of 79.5. The distribution shows that the results are negatively skewed with 2 outliers on the upper end of the scale.

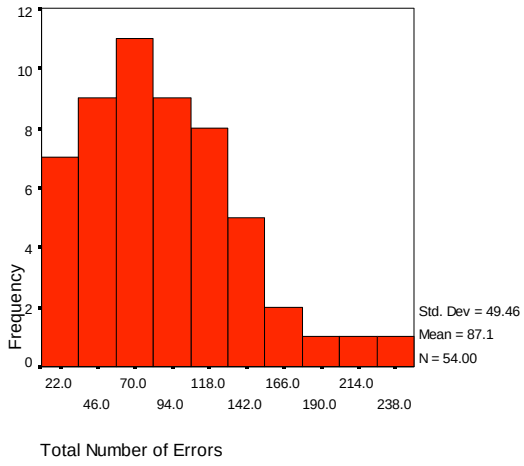

Figure 2. Total Number of Errors for All Tests

The results for the ANOVA and Scheffe test for the total number of errors over the three tests between the three groups of participants are shown in Table 1.

	Sum of Squares	df	Mean Square	F	P value
Between Groups	37451.704	2	18725.852	10.358	.000
Within Groups	92202.389	51	1807.890		
Total	129654.093	53			

Advanced (I), Beginner (J)		Advanced (I), Intermediate (J)		Intermediate (I), Beginner (J)	
Difference (I-J)	p value	Difference (I-J)	p value	Difference (I-J)	p value
-63.89	.000	-39.67	.026	-24.22	.242

Table 1. ANOVA and Scheffe of between participant groups for Total Number of Errors for advanced, intermediate and beginner participants of the Online Grocery System

For the Total Number of Errors the p value of 0.000 indicates that there is statistical significance of the difference at the 5% significance level. The results of the Scheffe comparison (see Table 1) indicate that there is a statistical significance at the 5% level between advanced and beginner participants (p value of 0.000), and between advanced and intermediate participants (p value of 0.026). However, there is no statistical significance between intermediate and beginner participants (p value of 0.242) at either the 5% or 10% level.

An analysis of the types of errors made, using the 'Taxonomy of Errors' (Zapf et al. 1992), is essential to identify ways of improving the Online Grocery Systems.

Usability Problems

Table 2 below lists the number of usability problems made by each participant type in each test.

	Short Test	Medium Test	Long Test	Total
Advanced Participants	71	141	296	508
Intermediate Participants	118	229	489	836
Beginner Participants	132	230	543	905
Total	321	600	1328	2249

Table 2. Total number of usability problems

Usability problems accounted for 47.8% (N=2249) of all of the errors observed. Many of these errors were only recorded once and some specific errors were only made by one participant. These errors will not be discussed in detail due to their low frequency. Errors that occurred multiple times and were made by multiple participants will be discussed in some detail.

Most of the participants had never interacted with an Online Grocery System prior to participating in this research. Of those that had used an Online Grocery System, few had used the system reviewed in this study. It is therefore expected that there will be a higher percentage of errors per product for the short test. This should be especially true for sensorimotor, knowledge base and intellectual regulation errors. Sensorimotor errors accounted for 31.7% of all the errors that participants encountered. This was the most common type of usability problem that participants faced. The total time that participants spent recovering from sensorimotor errors was 3658 seconds. The average time spent recovering per error was 2.45 seconds. The time taken to recover from a sensorimotor error, on average, was the lowest of all error types.

For the short and medium tests, beginner and intermediate participants made a similar number of sensorimotor errors. This similarity may be explained by the fact that both participant groups were inexperienced in their interactions with Online Grocery Systems. During the long test, advanced participants (as a group) made fewer sensorimotor errors than intermediate or beginner participants, but beginners made more errors than intermediate participants. This result suggests that the intermediate participants were able to learn the system more quickly than beginner participants, thus reducing their number of errors faster than beginners.

The number of sensorimotor errors, particularly those which resulted in the correct product not being returned due to spelling mistakes, could be reduced if the Online Grocery System under review had an inbuilt spelling function.

Inefficiency problems

Table 3 below lists the number of inefficiency problems experienced by each participant type in each test.

	Short Test	Medium Test	Long Test	Total
Advanced Participants	23	19	56	98
Intermediate Participants	77	149	229	455
Beginner Participants	109	158	417	684
Total	209	326	702	1237

Table 3. Total number of inefficiency problems

Inefficiency problems accounted for 26.3% of the errors observed. The average time a participant spent recovering from an inefficiency problem was 3.83 seconds. Most of the inefficiency problems were made by beginners. The number of inefficiency problems increased as the product list length was increased, with some participants repeating the same error for each product. Such errors included participants adding the quantity of one before clicking the checkbox, and using the pop-up box.

It would be difficult for the number of inefficiency problems to be decreased as many of the participants recorded a large number of errors due to habit in the long test. Participants' habit of using pop-up boxes was demonstrated when, for some products, a pop-up box was unavailable. In such cases, participants added the product in the most efficient way. However, on subsequent products with a pop-up box available, participants continued to add the product via the pop-up box.

Functionality Problems

Table 4 below lists the number of functionality problems experienced by each participant type in each test.

	Short Test	Medium Test	Long Test	Total
Advanced Participants	55	37	249	341
Intermediate Participants	50	47	273	370
Beginner Participants	73	85	350	508
Total	178	169	872	1219

Table 4. Total number of Functionality problems

Functionality problems accounted for 25.9% of the errors observed. The four types of functionality errors are discussed individually below. Action blockages accounted for 14% of all of the errors observed. All action blockage errors occurred because a product was not available at the time that the participant completed the tests. The average time spent recovering from action blockages 23.05 seconds. The recovery time was the longest of all of the errors. Action blockage errors could have been completely avoided if the Online Grocery System listed all products in their inventory, with the appropriate products identified as 'out of stock'. This practice is used by some Online Grocery Systems.

CONCLUSIONS

By definition, errors can be both attributed to actions of the system and the user. However, as an error depends on the attempted obtainment of a goal, a system cannot commit an error because it does not have any intentions of its own. Therefore an error only occurs when a user interacts with a system. From the viewpoint of the users this is not always the case, for instance a system crash or a bug in the system; the user usually believes that it is the system that has caused the error. In essence these errors can be attributed to either other users or humans. However, according to Zapf et al. (1992) this can only be attributed in a theoretical sense. From a scientific view point an error is always due to a mismatch between the user and the system.

The errors made on the Online Grocery System by all three user types were successfully classified using the 'Taxonomy of Errors', demonstrating that the taxonomy is appropriate for application on non-work related systems used by casual users. The taxonomy also provided an effective means of comparing different errors across different participant groups.

It is generally assumed that beginner users make more errors than intermediate and advanced users. Previous studies have found that this was not the case (Zapf et al. 1992; Hyland 2001), however beginner users did make different types of errors. This study confirmed that beginner users make different types of errors to more experienced users. This study is the first to demonstrate a situation where beginner users made significantly more errors than more experienced users.

REFERENCES

Ahuja, M., Gupta, B., and Raman, P. An Empirical Investigation of Online Consumer Purchasing Behavior. Communications of the ACM 46,12 (2003), 145-151.

Brodbeck, F.C., Zapf, D., Prumper, J., and Frese, M. Error handling in office work with computers: A field study. Journal of Occupational and Organizational Psychology 66 (1993), 303-317.

Cochran, W.G., and Cox, G.M. Experimental Designs. New York, Wiley (1957).

Harper, J., Caputi, P., Jayasiriya, R., Vella, S., and Hyland P. Using a Taxonomy of Errors as a Conceptual Framework for Differences in Patterns of Use for Casual and Novice Users. In Proc. OZCHI2004 (2004).

Hoc, J., Cacciabue, P.C., and Hollnagel, E. Editorial. Expertise and Technology: Cognition & Human-Computer Cooperation. J. Hoc, P.C. Cacciabue and E. Hollnagel Eds. Hillsdale, NJ, Lawrence Erlbaum Associates (1995), 1-14.

Hyland, P. Supporting the Retrieval and Analysis of External Statistical Data by Casual and Novice Users. Department of Information Systems. Wollongong, University of Wollongong (2001).

Kalin, S. Mased and Confused. WebBusiness Magazine. (1999).

Leadbetter, D., Hussey, A., and Lindsay, P. Towards Model Based Prediction of Human Error in Interactive Systems. IEEE (2001), 42-49.

Nielsen, J. Test With 5 Users, Alertbox, March 19. Retrieved from http://www.useit.com/alertbox/20000319.html (2000).

Nielsen, J. First Rule of Usability: Don't Listen to Users, Alertbox for 5 August 2001. Retrieved from http://www.useit.com/alertbox/20010805.html (2001).

Prümper, J., Zapf, D., Brodbeck, F.C., and Frese, M. Some Surprising Differences Between Novice and Expert Errors in Computerized Office Work. Behaviour and Information Technology 11, 6 (1992), 319-328.

Rasmussen, J. Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. New York, North-Holland (1986).

Rauterberg, M. Why and What can we Learn from Human Errors? Advances in Applied Ergonomics (1996), 827-830.

Rauterberg, M., and Felix, D. Human Errors: Disadvantages and Advantages. In Proc. Pan Pacific Conference on Occupational Ergonomics (1996).

Reason, J. Human Error. Cambridge [England], Cambridge University Press (1990).

Rubin, J. Handbook of Usability Testing: how to plan, design and conduct effective tests. New York, Wiley (1994).

Sutcliffe, A., and Rugg, G. A Taxonomy of Error Types for Failure Analysis and Risk Assessment. International Journal of Human-Computer Interaction 10, 4 (1998), 381-405.

Wickens, C.D., and Hollands, J.G. Engineering Psychology and Human Performance 3rd Edition. Upper Saddle River, NJ, Prentice Hall (2000).

Zapf, D., Brodbeck, F.C., Frese, M., Peters, H., and Prümper, J. Errors in Working with Office Computers: A First Validation of a Taxonomy for Observed Errors in a Field Setting. International Journal of Human-Computer Interaction 4, 4 (1992), 311-339.